

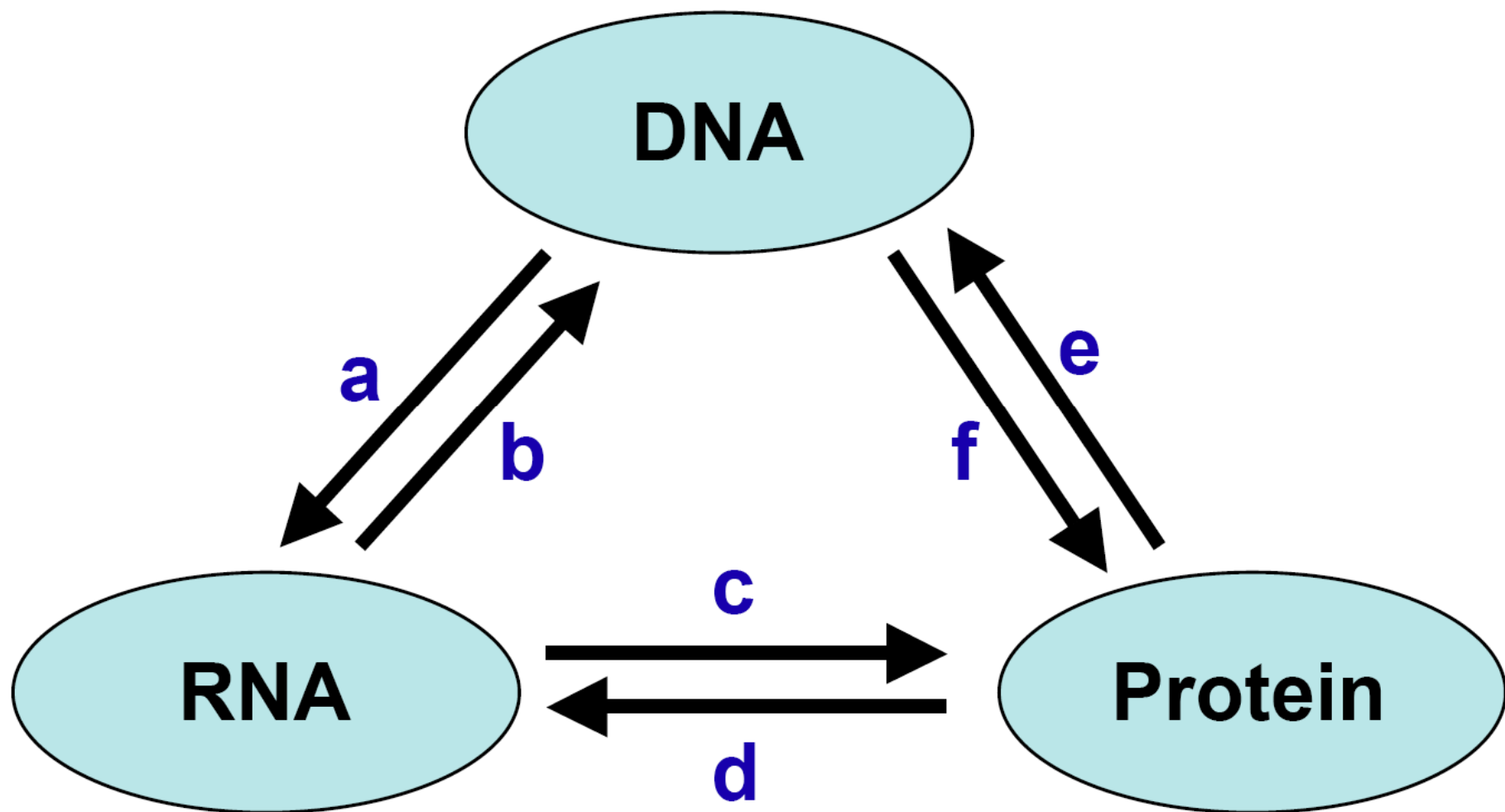
# DNA as Biological Information

Rasmus Wernersson

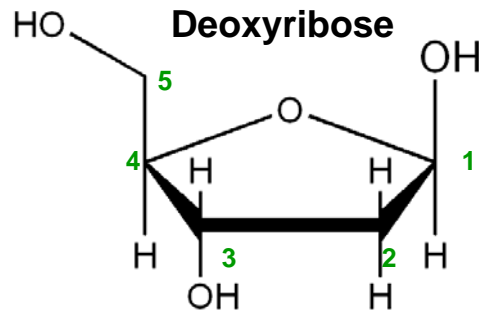
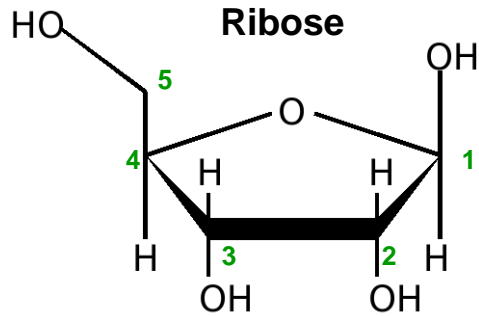
Henrik Nielsen

- Learning objectives
  - About Biological Information
  - A note about DNA sequencing techniques and DNA data
  - File formats used for biological data
  - Introduction to the GenBank database

# Information flow in biological systems



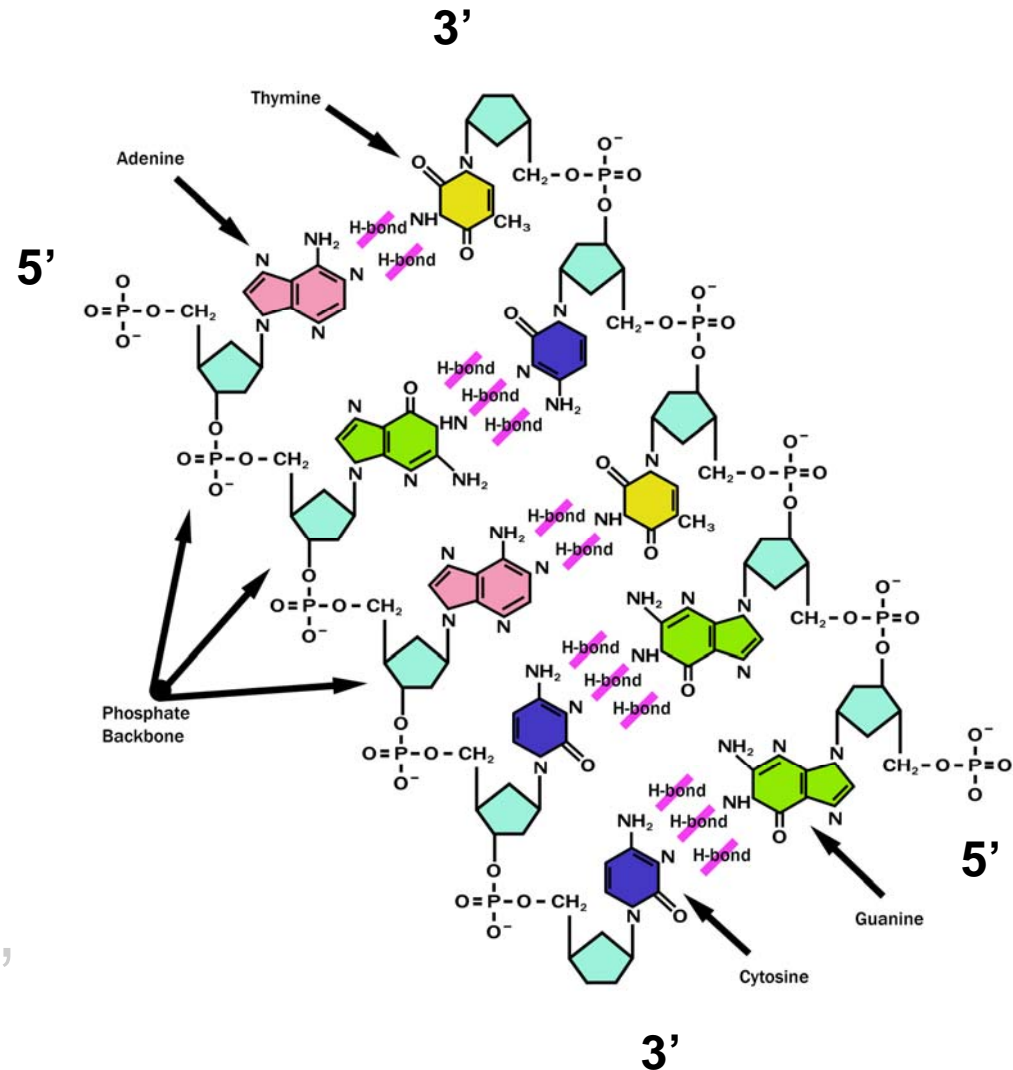
# DNA sequences = summary of information



5' AGCC 3'

3' TCGG 5'

5' ATGGCCAGGTAA 3'



DNA backbone: <http://en.wikipedia.org/wiki/DNA>

(Deoxy)ribose: <http://en.wikipedia.org/>

Cycle 1

5'-CTAGGATATGAAACCTATAGGTACGGTGGCCATTCTATGTCTGATCCCGGTACTACCTACAGAA-3'

|||||  
 3'-GGCCATGATGG-5'

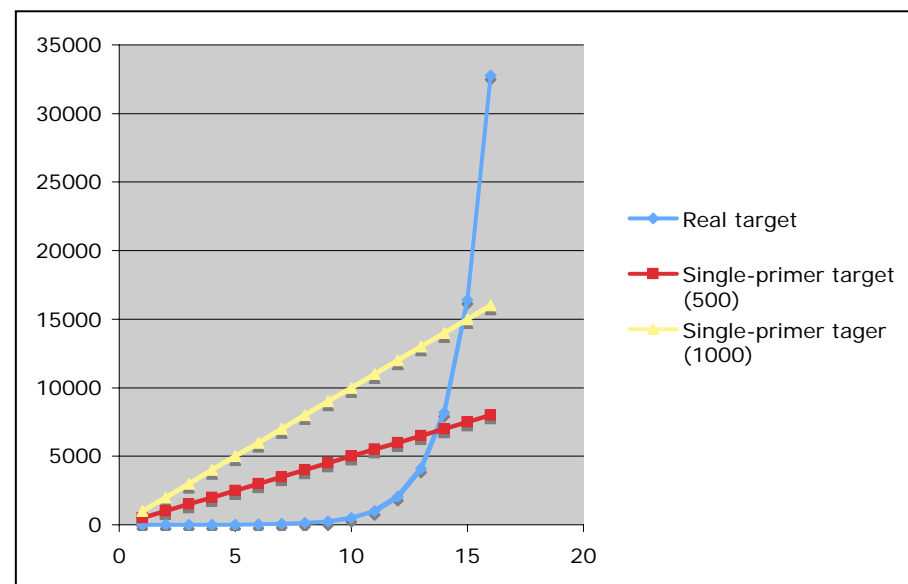
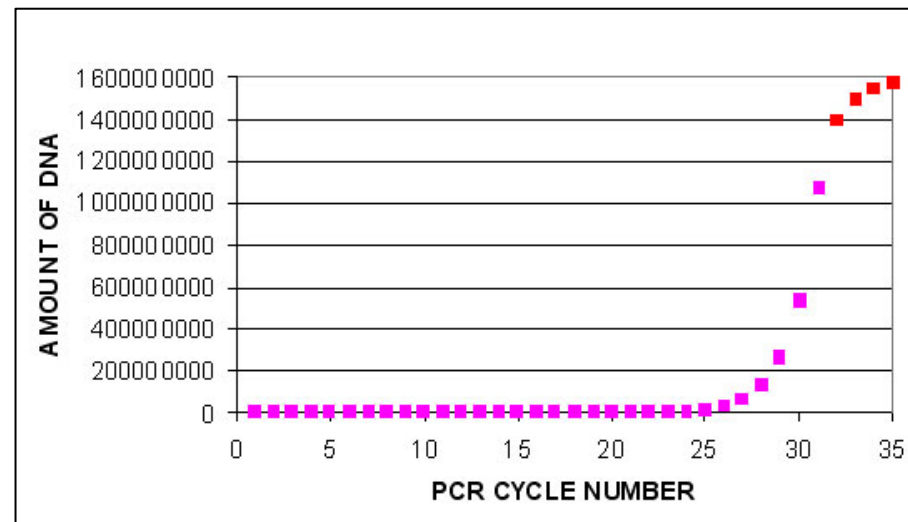
35  
 cycles

**Melting**  
 96°, 30 sec

**Annealing**  
 ~55°, 30 sec

**Extension**  
 72°, 30 sec

5'-ATGAAACCTATAG-3'  
 |||||  
 3'-GATCTTATACTTTGGATATCCATGCCACCGGTAAGATACAGACTAGGGCCATGATGGATGTCTT-5'

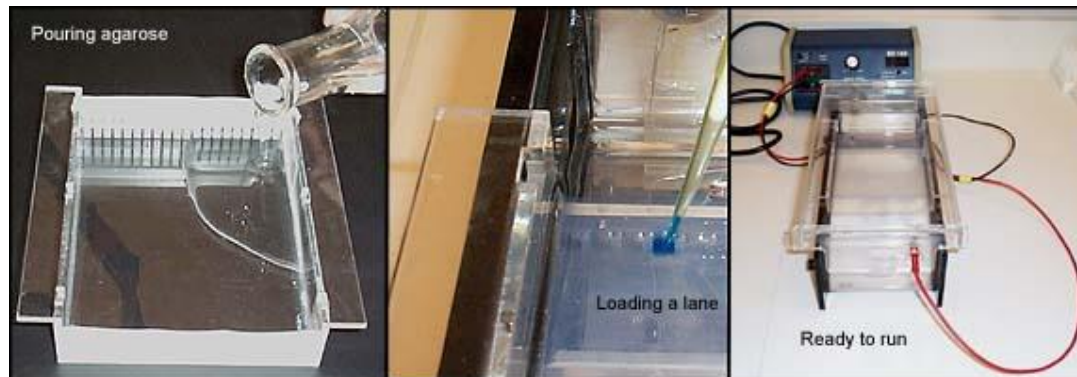
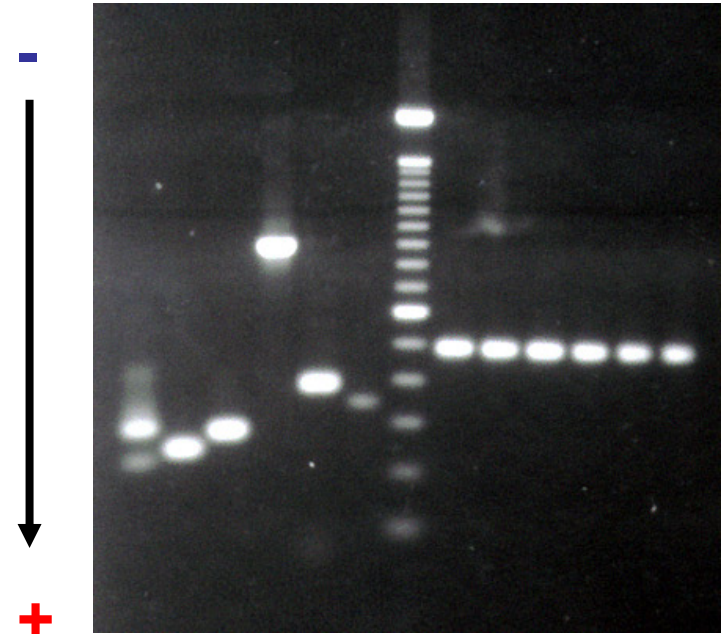


**Animation:** <http://www.people.virginia.edu/~rjh9u/pcranim.html>

**PCR graph:** <http://pathmicro.med.sc.edu/pcr/realtime-home.htm>

# Gel electrophoresis

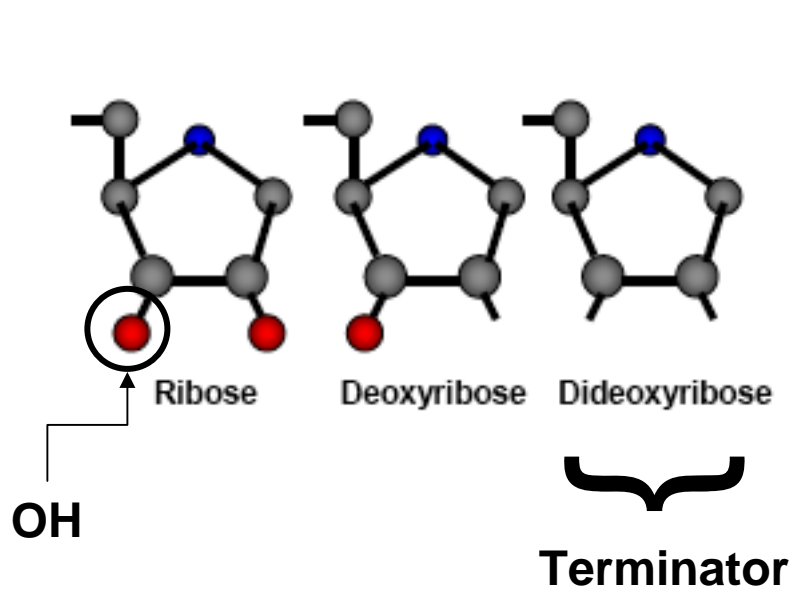
- DNA fragments are separated using gel electrophoresis
  - Typically 1% agarose
  - Colored with EtBr or ZybrGreen (glows in UV light).
  - A DNA "ladder" is used for identification of known DNA lengths.



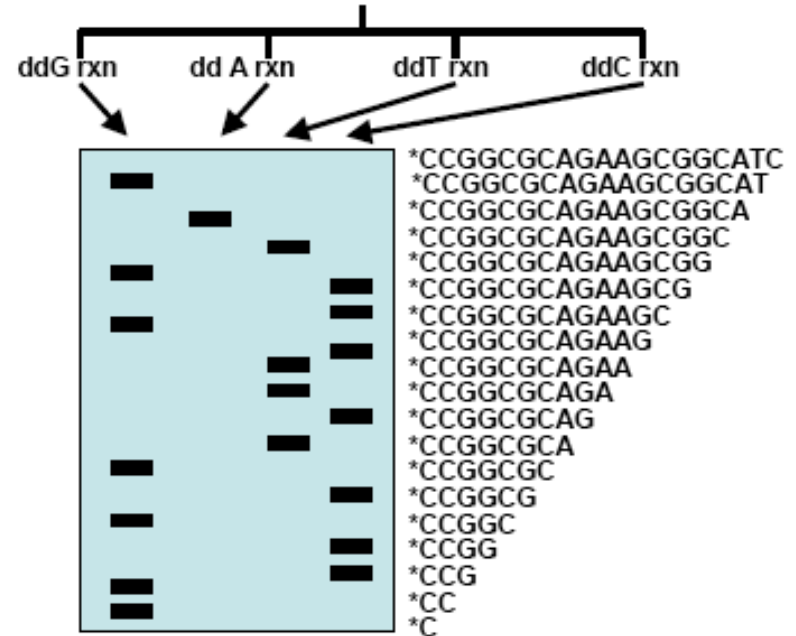
Gel picture: <http://www.pharmaceutical-technology.com/projects/roche/images/roche3.jpg>

PCR setup: <http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/gels/agardna.html>

# The Sanger method of DNA sequencing



5' pCpCpGpGpCpGpCpApGpApApGpCpGpGpCpApTpCpApGpCpApApA 3'

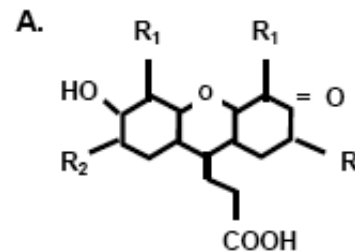


X-ray sequencing gel



## Automated sequencing

- The major break-through of sequencing has happened through *automation*.
- Fluorescent dyes.
- Laser based scanning.
- Capillary electrophoresis
- Computer based base-calling and assembly.

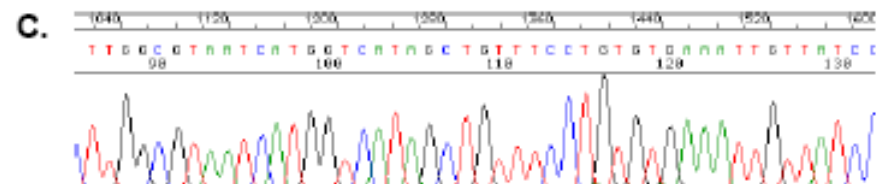
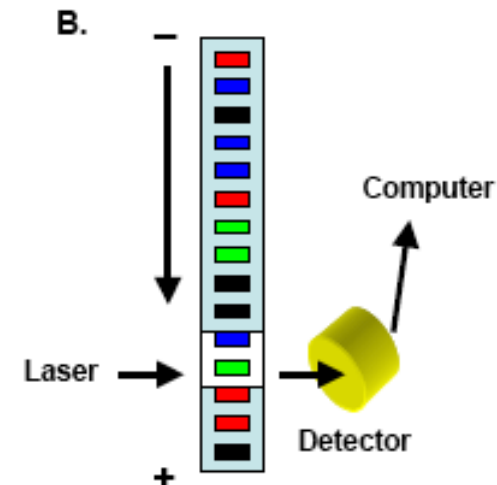
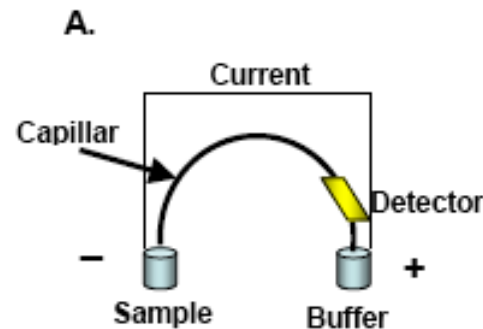
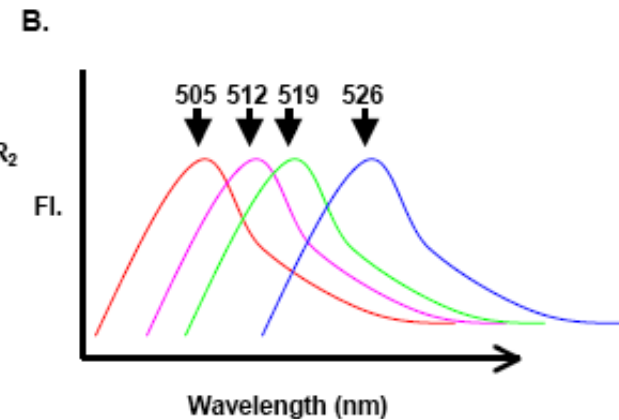


SF505:  $R_1=R_2=H$

**SF512:**  $R_1=H$ ,  $R_2=CH_3$

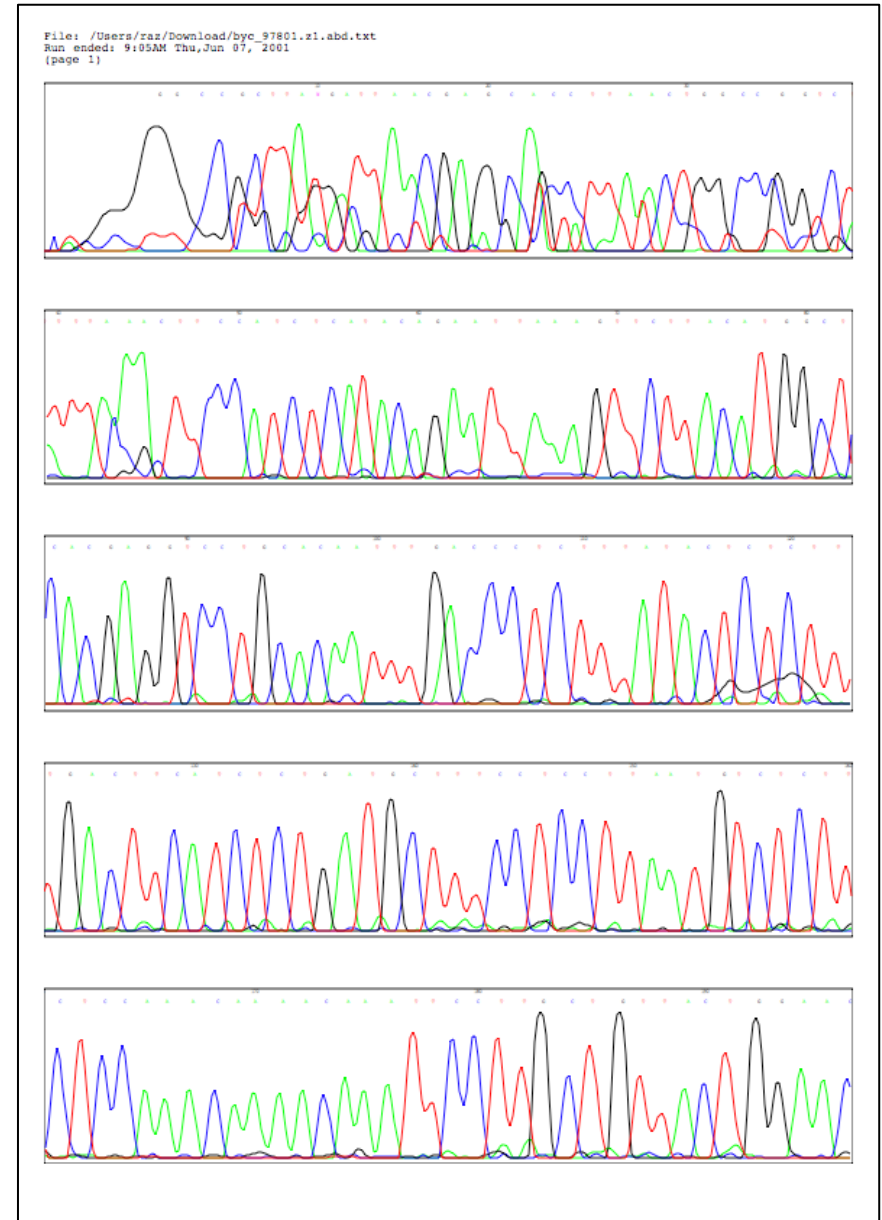
SF519:  $R_1=CH_3$ ,  $R_2=H$

**SF526:**  $R_1=R_2=CH_3$



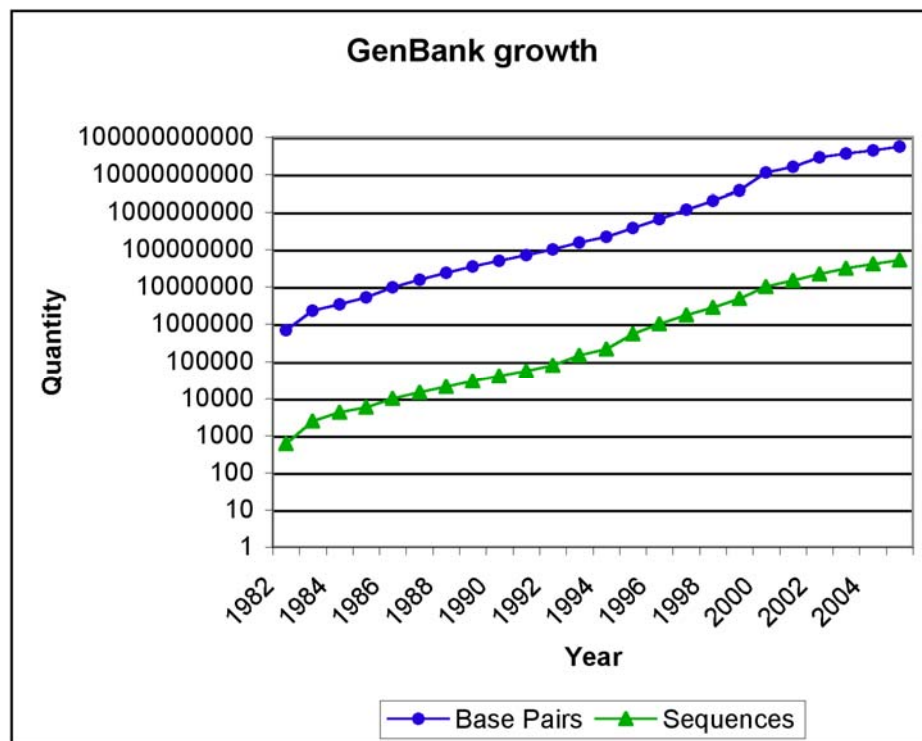
# Handout exercise: "base-calling"

- Handout:  
Chromotogram
- Groups of 2-3.
- Tasks:
  - Identify "difficult" regions
  - Identify "difficult" sequence stretches.
  - Try to estimate the best interval to use.



- The GenBank database
- File formats
  - FASTA
  - GenBank

- GenBank is one of the main international DNA databases.
- GenBank is hosted by NCBI: *National Center for Biotechnology Information.*
- GenBank has existed since 1982.
- The database is public - no restrictions on the use of the data within.



## >alpha-D

```
ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCAC
CCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCA
AGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTGTTTCGGCTGGGGGTCAG
CACTGACCATCCCGCTCCCGCAGCTGTTTACCACCTACCCCCAGACCAAGACCTACTTCC
CCCCTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGG
CCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA
GCGACCTGCATGCCTACAACCTGCGTGTGACCCCTGTCAACTTCAAGGCAGGCGGGGGAC
GGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCC
GGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTG
GCCACACACCTGGGCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTG
TCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA
```

## >alpha-A

```
ATGGTGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAATCGGCGGC
CAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCC
ATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCATCCATAACTG
TCCCTGTTCTATGTGGCCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTC
TGTCCCCCAGGTTGTTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACC
TGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTG
AGGCTGCCAACCACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACG
CCCAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCA
GTCTGGCTCCCCCTCCTGCACACACCTCTGGCTACCCCTCACCTCACCCCTTGCTCACC
ATCTCCTTTTGCCTTTTCACTGTCTGGGTCACTGCTTCCCTGGTGGTCGTGGCCGTCCACTT
CCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGG
CACCGTCCTTACTGCCAAGTACCGTTAA
```

## GenBank format

[illegible]

## Header

Indeholder  
information ang.  
Organisme,  
publikation,  
Accession ID mm.

## FEATURE blok

Indeholder en beskrivelse af forskellige elementer i DNA sekvensen.

**CDS: Coding Sequence.**  
Indeholder koordinater på den protein kodende del af et gen. Bemærk de tre intervaller.

ORIGIN blok

Indeholder selve  
DNA sekvensen.

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

LOCUS CMGLOAD 1185 bp DNA linear VRT 18-APR-2005  
 DEFINITION Cairina moschata (duck) gene for alpha-D globin.  
 ACCESSION X01831  
 VERSION X01831.1 GI:62724  
 KEYWORDS alpha-globin; globin.  
 SOURCE Cairina moschata (Muscovy duck)  
 ORGANISM Cairina moschata  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.  
 REFERENCE 1 (bases 1 to 1185)  
 AUTHORS Erbil,C. and Niessing,J.  
 TITLE The primary structure of the duck alpha D-globin gene: an unusual  
 5' splice junction sequence  
 JOURNAL EMBO J. 2 (8), 1339-1343 (1983)  
 PUBMED 10872328  
 COMMENT Data kindly reviewed (13-NOV-1985) by J. Niessing.

# GenBank format - ORIGIN section

## ORIGIN

```

1  ctgcggtggcc tcagccctc caccctcca cgctgataag ataaggccag ggcgggagcg
61  cagggtgcta taagagctcg gcccgcggg tgtctccacc acagaaaccc gtcagttgcc
121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
241 gctgggcca gggggcactc acagggtggg cagcaggagg caggagccct gcagcgggtg
301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
361 aaaactgact ggcctcgctc cggcaggatg ttctcgcct acccccagac caagacctac
421 ttccccact tcgacctgca tcccggtctt gaacagggtc gtggccatgg caagaaagtg
481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
541 ctgagcaacc tgcattgcta caacctgcgt gttgaccctg tcaacttcaa ggcaagcggg
601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcagggtctg ggggtccagg
661 ggtctgagtt tctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
721 gggtagcagg gtcttggggg ccagcagcca gacagcaggg gctgggattg catctgggat
781 gtgggcccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
901 tccggagcag gggtagtaag ccctggtttg ccttgcagct gctggcacag tgcttccagg
961 tgggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
1021 agttcttgtc cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc

```

//



# GenBank format - FEATURE section

```
FEATURES                     Location/Qualifiers
    source                    1..1185
                              /organism="Cairina moschata"
                              /mol_type="genomic DNA"
                              /db_xref="taxon:8855"
    CAAT_signal               20..24
    TATA_signal               69..73
    precursor_RNA             101..1114
                              /note="primary transcript"
    exon                       101..234
                              /number=1
    CDS                       join(143..234,387..591,939..1067)
                              /codon_start=1
                              /product="alpha D-globin"
                              /protein_id="CAA25966.2"
                              /db_xref="GI:4455876"
                              /db_xref="GOA:P02003"
                              /db_xref="InterPro:IPR000971"
                              /db_xref="InterPro:IPR002338"
                              /db_xref="InterPro:IPR002340"
                              /db_xref="InterPro:IPR009050"
                              /db_xref="UniProt/Swiss-Prot:P02003"
                              /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
QCFQVVLAAHLGKDYSPMHAAFDKFLSAVAAVLAEKYR"
    repeat_region             227..246
                              /note="direct repeat 1"
    intron                    235..386
                              /number=1
    repeat_region             289..309
                              /note="direct repeat 1"
    exon                       387..591
                              /number=2
    intron                    592..939
                              /number=2
    exon                       940..1114
                              /number=3
    polyA_signal              1095..1100
    polyA_signal              1114
```

# Exercise: GenBank

- Work in groups of 2-3 people.
- The exercise guide is linked from the course programme.
- Read the guide carefully - it contains a lot of information about GenBank.

The screenshot shows the Entrez Nucleotide database homepage. The browser address bar displays the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=>. The page features a navigation bar with links to various databases: All Databases, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. A search bar is prominently displayed with the text "Search Nucleotide for" and buttons for "Go" and "Clear". Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area includes a paragraph about the Entrez Nucleotides database, a section titled "Human Genome" with a link to "human genome resources", and a section titled "Building the human genome" with a link to "Genome View". At the bottom, there is a "Homo sapiens genome view" section showing a bar chart of genome statistics.

Entrez Nucleotide

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=>

Overview - p...oc.perl.org IT-Diplomud...elsen (ITD) Password generator Nucl. Acids R...rver Issue

NCBI Nucleotide

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for

Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Entrez Nucleotide Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI

Related resources BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of June 2005, there are over 89 billion bases in GenBank and RefSeq alone.

Human Genome

Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

Building the human genome

The Human Genome Reference DNA Sequence was completed in April 2003. The current version is listed as a build number on the [Genome View](#) page and includes an accompanying set of [statistics](#) and [release notes](#).

Homo sapiens genome view

build 35 version 1 statistics

Hit

1 2 3 4 5 6 7 8 9 10 11 12 13

14 15 16 17 18 19 20 21 22 X Y

Go to "<http://www.ncbi.nlm.nih.gov/FLC/getmgc.cgi>"